ABSTRACT

The establishment of Semantic Similarity between input text units forms an integral part of several Natural Language Processing applications. Paraphrasing is a specific form of similarity wherein the input units exhibit semantic equivalence but may have considerable differences in terms of the lexical units as well as syntactic structures. Detection of paraphrases is crucial to the success of applications such as multi-document summarization, Question Answering and Plagiarism detection. The task is complicated due to the variability and ambiguity of natural language inputs. The wide applicability as well as the challenging nature of the task has inspired active research in this domain by both academic and industry research groups. This research focuses on the design of machine learning based systems for two core areas related to sentence-level Paraphrasing, namely Recognition and Extraction.

Existing systems for Paraphrase Recognition employ a variety of methods including vector space models, machine learning, rule decoding, graph matching and logic based approaches. Due to the complexity of the task, there is a need for developing efficient systems. In this thesis, four different machine learning based approaches have been adopted for sentence level paraphrase recognition. In the first approach, a Radial Basis Function Neural Network (RBFNN) has been designed to operate on lexical, syntactic and semantic features extracted directly from the input sentences to identify paraphrases. The use of Neural Networks has been explored due to their good generalization capability. Orthogonal Least Squares Learning method has been used to construct the neural network. The approach has been evaluated on two standard paraphrase corpora - Microsoft Research Paraphrase Corpus (MSRPC) and extended Knight & Marcu Corpus (KMC). The RBFNN based system has exhibited marginal improvement over existing Support Vector

Machine (SVM) Paraphrase Recognizers while incurring additional training time. The feature set used was found to greatly influence the performance of the recognition system. Therefore an SVM based system for Paraphrase Recognition was designed by optimizing the input feature set. Wrapper-based Feature selection using Genetic Algorithms was employed to identify the best set of features. When evaluated on the MSRPC and extended KMC, the system was found to yield better performance with the added advantage of using only half the original number of features. Of the various categories of features, lexical features were found to perform best. Including additional inputs in the form of a table of equivalent phrases was found to enhance the performance further.

Translation of the input sentences to an intermediate representation enables easier matching and is therefore expected to simplify the Paraphrase Recognition process. Two such intermediate representations have been explored in this research. Universal Networking Language (UNL) constitutes the first intermediate representation and has the advantage of being applicable to cross-language inputs. In the UNL based Paraphrase Recognition (UNLPR) system, the UNL components of the two sentences were matched to generate features. These were then used by a Support Vector Machine classifier to detect paraphrase pairs. Since this method ignores word overlap as well as syntactic features it has exhibited the lowest performance when compared to the three other approaches. The second intermediate representation employed was Predicate Argument Structures (PAS) which convey semantic roles and therefore enable deeper matching of the input text. The PAS based Paraphrase Recognition system (PASPR) works by first pairing the predicate-argument tuples and segregating the sentences based on the presence of paired, loosely paired and unpaired tuples. Surface-level features extracted from each category have then been used to recognize the paraphrases. This system was found to outperform all the other proposed systems as well as the current best

performing Paraphrase Recognition system due to matching of semantic roles followed by the usage of category-specific features.

Paraphrase Extraction refers to the discovery of paraphrases from a large scale corpus and serves as the second focus area of this research. Fuzzy Hierarchical Clustering approach has been used to design a sentence-level paraphrase extraction system. Sentences describing the same actions have been grouped using Fuzzy agglomerative clustering based on verbs. In the next step, fuzzy divisive clustering centered on nouns was applied to refine the clusters followed by the deployment of paraphrase recognizers to identify the paraphrases within each cluster. The system has exhibited promising results on a subset of the Microsoft Research Video Description Corpus as well as the MSRPC, when compared to traditional hard and soft clustering schemes. The reason for the better performance of the Fuzzy Hierarchical Clustering technique can be attributed to the ability to create adaptive and overlapping clusters.

The proposed Paraphrase Recognition and Extraction systems have been deployed in various practical applications. A Short Answer Evaluation system which considers correct student answers as paraphrases of the reference answer has been developed. A Plagiarism detection system has also been designed to handle sophisticated forms of plagiarism such as paraphrasing. Experimental evaluation using standard short answer corpora and plagiarism corpora have proved the effectiveness of using Paraphrase Recognition approach. A scheme for clustering news headlines has been developed by using the proposed Paraphrase Extraction system and was found to perform satisfactorily on a sample collection of Google news headlines. The Paraphrase Recognition and Extraction systems designed in this research are found to have wide applicability in real world systems.